LPENS
LABORATOIRE DE PHYSIQUE
DE L'ÉCOLE NORMALE SUPÉRIEURE

SORBONNE
UNIVERSITÉ

Machine-learning-assisted Monte Carlo fails at sampling computationally hard problems

Simone Ciarella, **Jeanne Trinquier**, Martin Weigt, Francesco Zamponi

arXiv:2210.11145

Motivations

- Goal: generate equilibrium samples from the Boltzmann distribution $P_B(\sigma) = \frac{e^{-\beta H(\sigma)}}{Z}$
- Compute macroscopic properties of the system
- Combinatorial optimization (zero T limit)
- Universal strategy: local MCMC

MCMC

$$\mathrm{Acc}(\sigma_1, \sigma_2) = min(1, \frac{p_{\mathrm{Bo}}(\sigma_2) q(\sigma_2, \sigma_1)}{p_{\mathrm{Bo}}(\sigma_1) q(\sigma_1, \sigma_2)})$$

- Usually, **local** moves
- But there are hard-to-sample problems: high probabilities regions separated by low proba regions
- Sampling time can be exponential in the system size
- There exists system-specific solutions
- Recent line of research: machine learning assisted MCMC moves

Machine learning assisted MCMC

$$\mathrm{Acc}\left[\sigma_{old} \to \sigma_{new}\right] = \min\left[1, \frac{P_B(\sigma_{new}) \times P_{AR}(\sigma_{old})}{P_B(\sigma_{old}) \times P_{AR}(\sigma_{new})}\right]$$

- Suppose we have a generative model $P_{AR}(\sigma)$
- We can choose it as the transition kernel
- Global move: the proposed configuration is independent from the old one
- If the acceptance rate is high, very fast decorrelation

McNaughton, Milosevic, Perali, Pilati (2020)
Gabrié, Rotskoff, Vanden-Ejiden (2021)

An universal solution?

- Suppose we can find an ‘auxiliary’ distribution $P_{AR}(\sigma)$ such that
- It approximates well the target Boltzmann distribution

$$D_{KL}(P_B||P_{AR}) = \left\langle \log \frac{P_B(\sigma)}{P_{AR}(\sigma)} \right\rangle_{P_B} \ll N$$

Merchan, Nemenman (2016)

- It can be sampled easily e.g via autoregressive structure

$$P_{AR}(\sigma) = P^1_{AR}(\sigma_1)\, P^2_{AR}(\sigma_2|\sigma_1) \cdots P^N_{AR}(\sigma_N|\sigma_{N-1}, \cdots, \sigma_1)$$

Wu, Wang, Zhang (2019)

Learning the autoregressive distribution

- Standard approach to learn a generative model: maximum likelihood
- Minimize Kullback-Leibler divergence $D_{KL}(P_B||P_{AR}) = \left\langle \log \frac{P_B(\sigma)}{P_{AR}(\sigma)} \right\rangle_{P_B}$
- But it requires sampling from P_B
- Only possible at high T, allows to check the model expressivity

Maximum likelihood

If the model is expressive enough and the sample representative of the true landscape, we can reproduce the landscape

Only possible when data available

Data free

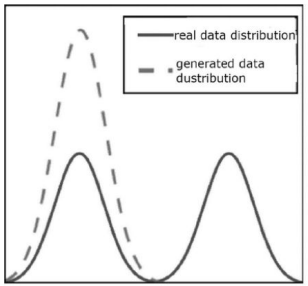
Might miss some parts of the landscape due to mode collapse

Variational approach

- Instead, minimize

$$D_{KL}(P_{AR}||P_B) = \left\langle \log \frac{P_{AR}(\sigma)}{P_B(\sigma)} \right\rangle_{P_{AR}} = \beta\,(F[P_{AR}] - F[P_B])$$

$$\beta \nabla_\theta F[P_{AR}] = \langle Q(\sigma) \nabla_\theta \log P_{AR}(\sigma) \rangle_{P_{AR}}$$
$$Q(\sigma) = \beta H(\sigma) + \log P_{AR}(\sigma)\ .$$



- To stop the learning, threshold on the variance of Q
- If the 2 distributions are the same: zero variance
- Reciprocally, if the variance is zero, then the 2 are equal but not necessarily on all configurations
- **Mode Collapse** (can be partially solved using simulated annealing)

Wu, Wang, Zhang (2019)

Simulated tempering

- Generate a sample via local MCMC at high T where sampling is easy
- Learn a first autoregressive model by maximum likelihood
- Decrease T and create a new sample with global MCMC

$$\mathrm{Acc}\left[\sigma_{old} \to \sigma_{new}\right] = \min\left[1, \frac{e^{-(\beta+\delta\beta)H(\sigma_{new})} P_{AR}(\sigma_{old})}{e^{-(\beta+\delta\beta)H(\sigma_{old})} P_{AR}(\sigma_{new})}\right]$$

McNaughton, Milosevic, Perali, Pilati (2020)
Gabrié, Rotskoff, Vanden-Ejiden (2021)

Autoregressive architectures

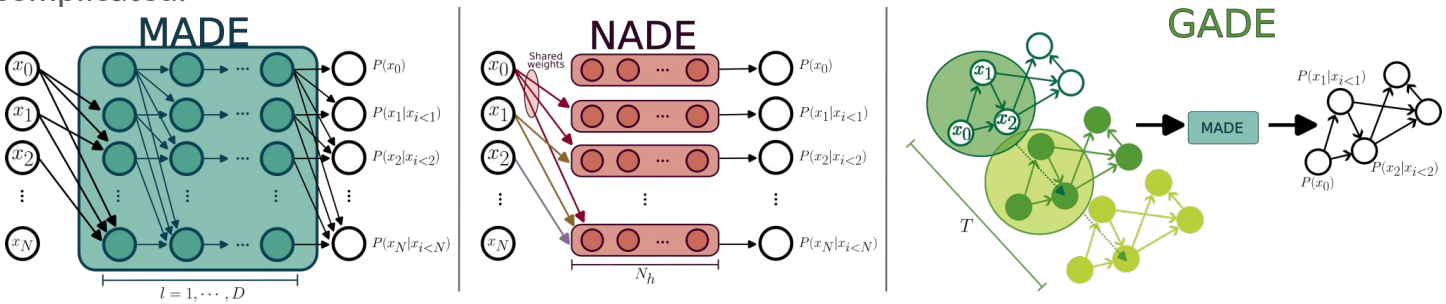
$$P_{AR}(\sigma) = P^1_{AR}(\sigma_1)\, P^2_{AR}(\sigma_2 \,|\, \sigma_1) \;\cdots\; P^N_{AR}(\sigma_N \,|\, \sigma_{N-1}, \cdots, \sigma_1)$$

Shallow MADE

$$P^1_{AR}(\sigma_1) = \frac{e^{h_1 \sigma_1}}{2\cosh(h_1)}$$

$$P^i_{AR}(\sigma_i \,|\, \sigma_{<i}) = \frac{e^{\sum_{j(<i)} J_{ij}\sigma_i\sigma_j + h_i\sigma_i}}{2\cosh\left(\sum_{j(<i)} J_{ij}\sigma_j + h_i\right)}$$

More complicated:



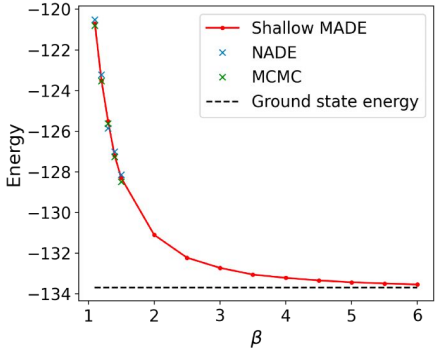
McNaughton, Milosevic, Perali, Pilati (2020)

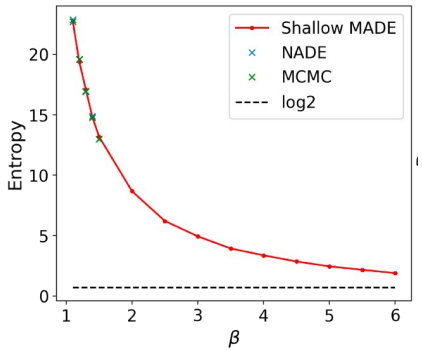
2d Edwards-Anderson spin glass

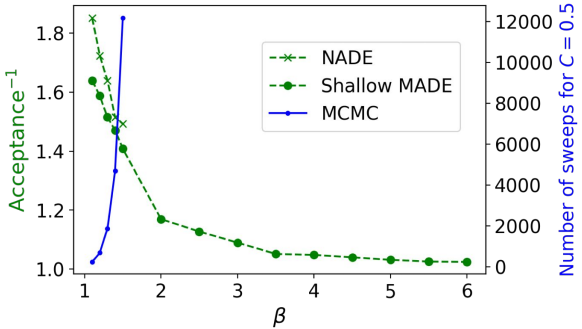
$$H(\sigma) = -\sum_{\langle i,j \rangle} J_{ij}\sigma_i\sigma_j \qquad \sigma_i \in \{-1,1\}$$

- Relaxation time of local MCMC grows fast when T is decreased
- Autoregressive models are able to recapitulate the energy and entropy
- High acceptance rate: decorrelation in a few steps

Local MCMC time /sweep	10^{-5} s
Global MCMC time/sweep	3 10^{-4} s
Training time at fixed β	5 mn

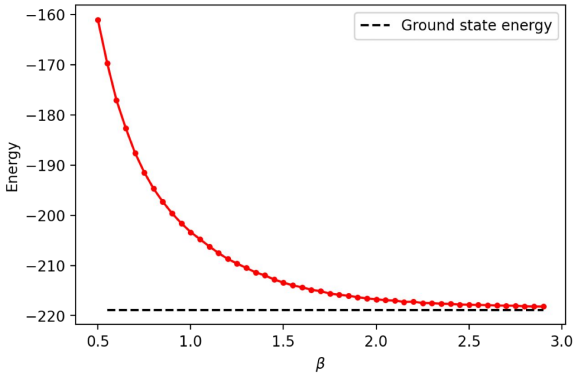


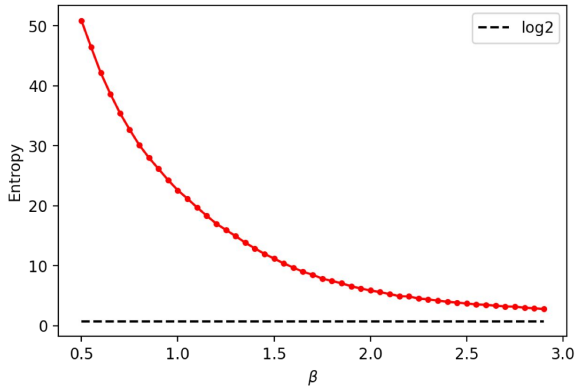


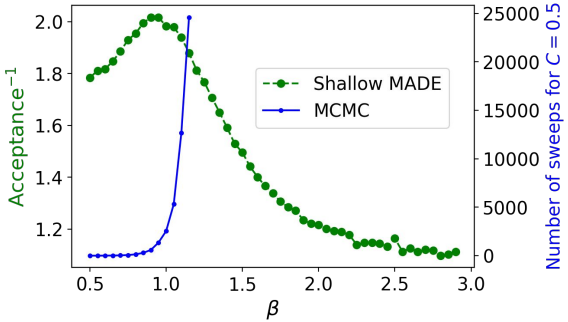


3d Edwards-Anderson spin glass

- Similar results despite the presence of a phase transition
- Needs a more systematic study on the dependance with the system size

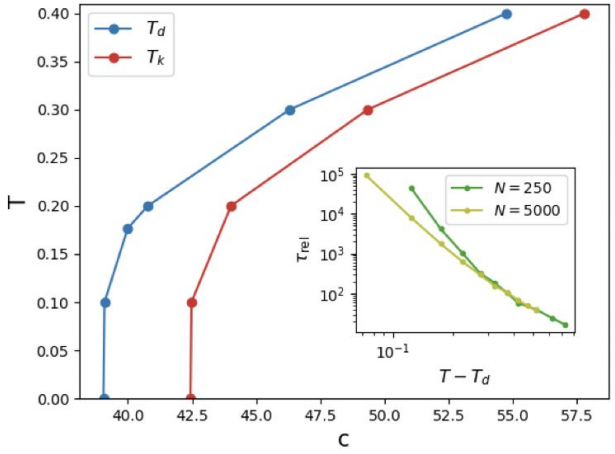






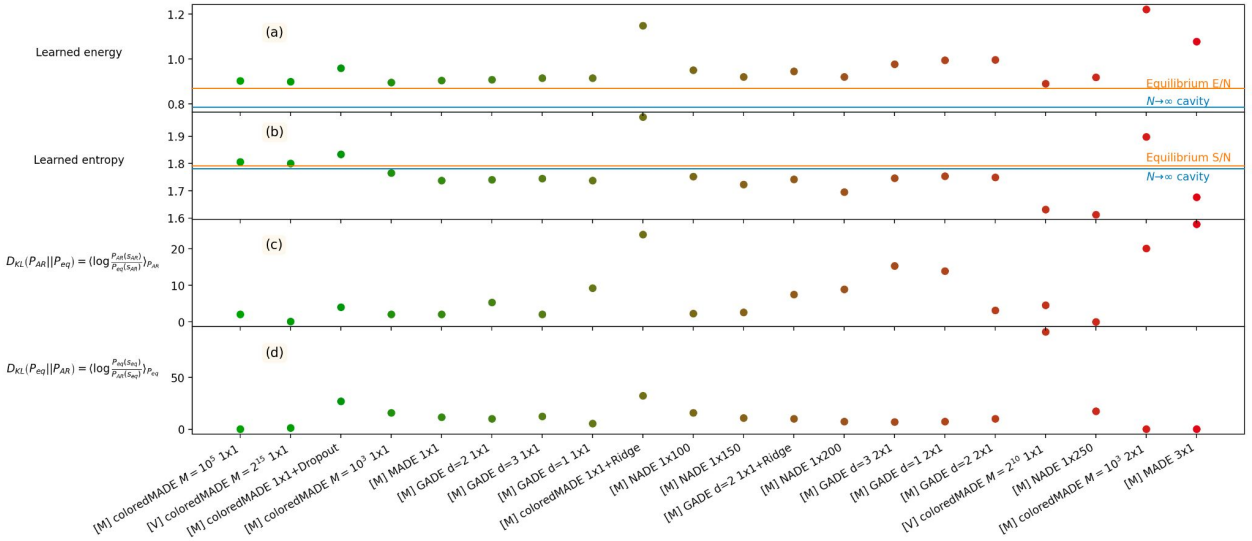
Failure for more complicated models $H(\sigma) = \sum_{\langle i,j \rangle \in \mathcal{G}} \delta_{\sigma_i,\sigma_j} \qquad \sigma_i \in \{1,2,\cdots,q\}$

- The sampling time of local MCMC is $(T\text{-}T_d)^{-\gamma}$ above T_d and $exp(N)$ below
- Erdos-Renyi graph with c = 40, q = 10
- System size N = 250
- Quiet planting can be used for any T to prepare one equilibrium configuration for a graph G



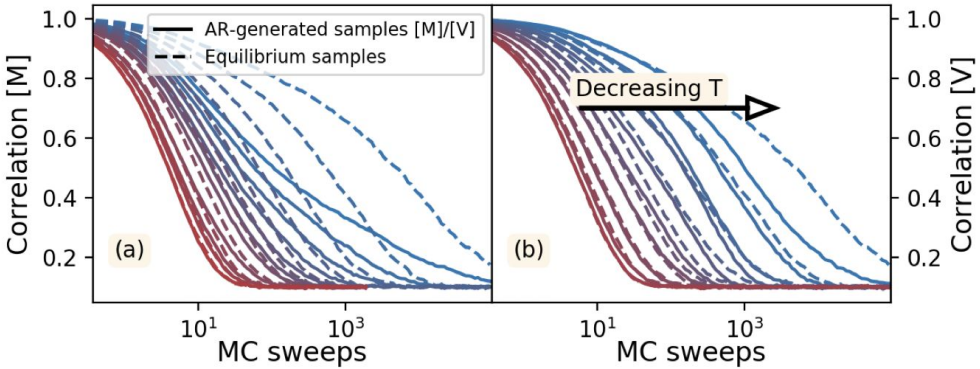
Krzakala, Zdeborová (2008)

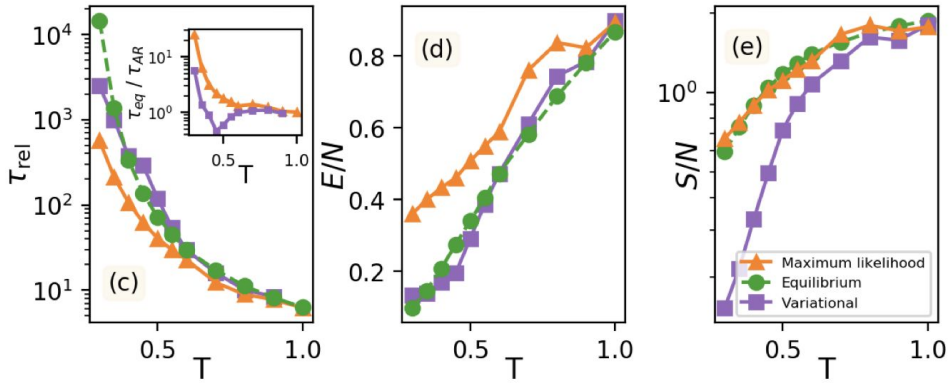
Model selection



- Learn models via maximum likelihood at [OBJ] where local MCMC is fast
- Increasing model expressivity leads to overfitting (too low entropy)
- Regularization leads to underfitting (too high entropy)
- Shallow MADE seems to be the best choice… [M] = Maximum likelihood, [V] = Variational

Lowering the temperature



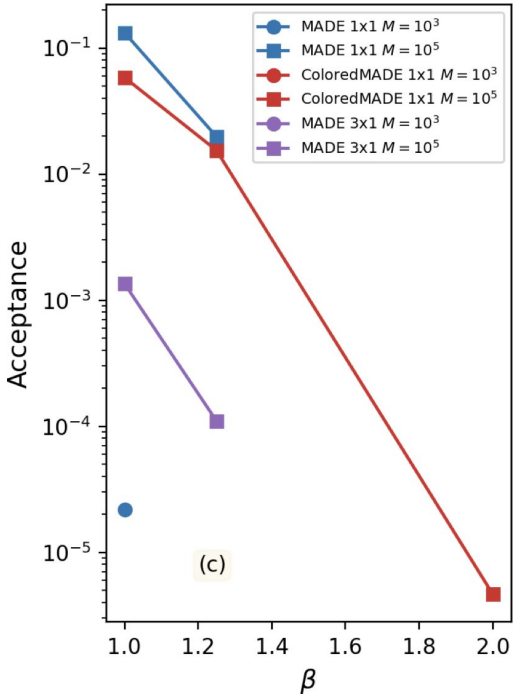


Local MCMC dynamics

[M] keeps good entropy, but too high energy

[V] has too low entropy, mode collapse

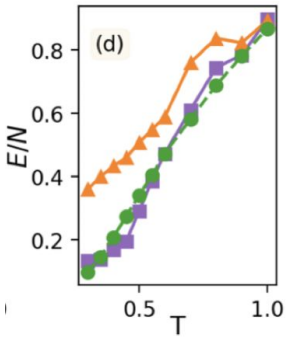
Maximum likelihood failure



- Recall: max-likelihood is the best we can do
- Using large enough training set is better…

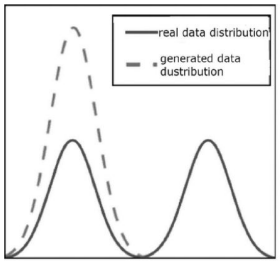
… but not enough,

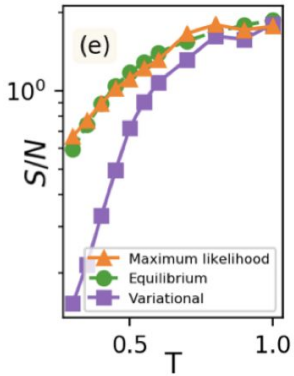
- The energy of the proposed configurations is too high

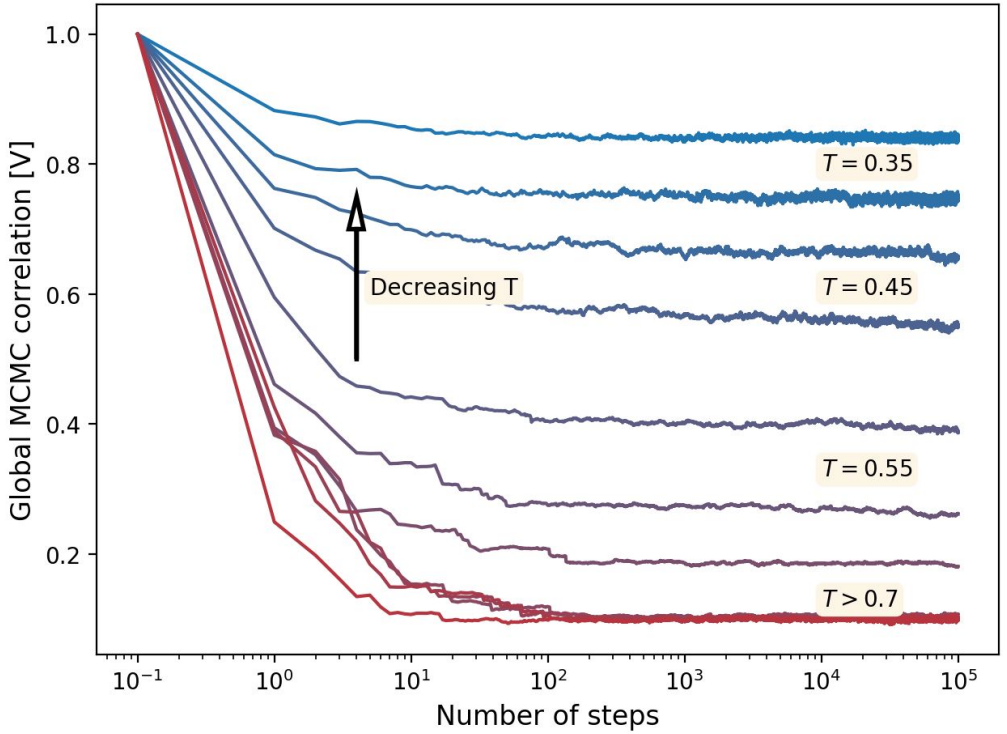


Variational failure

- Global MCMC dynamics, no decorrelation
- Moves are accepted, but do not lead anywhere







Perspectives

- Autoregressive models do not seem to provide a good enough auxiliary distribution for sampling complex models
- No clear path for improvement:
- More expressive models lead to overfitting
- Maybe increasing the training set could help, but need very large , computationally heavy
- Regularization leads to underfitting
- **Intrinsic limitation of the autoregressive structure ?** Can learn a few peaks, but not exp(N) peaks

Simplify the problem:
Learn to generate a subset given its boundaries

More complex architectures: transformers etc

- Reformulation of the Boltzmann distribution into an autoregressive one for CW and SK models: outperforms naive architectures
- System specific

Biazzo, 2023